Reinforcement Learning
Exercise 6

October 13, 2022

In this exercise we will implement two actor-critic reinforcement learning algorithms.

Policy gradient with a critic

Task 1 — 20 points Revisit the policy gradient solution for the InvertedPendulum from Exercise 5 with learned sigma and implement the actor-critic algorithm in `pg_ac.py`. Perform TD(0) updates at the end of each episode. **Attach the training performance plot into your report.**

Hint: Check out the PyTorch tutorial to see how to calculate the $A_\theta \nabla_\theta \log \pi_\theta(a_i|s_i)$ term using the `detach()` function.

The reference training plot is as Figure 1:

Question 1.1 — 10 points What is the relationship between actor-critic and REINFORCE with baseline?

Question 1.2 — 5 points How can the value of advantage be intuitively interpreted?

Question 1.3 — 10 points How does the implemented actor-critic method compare to REINFORCE in terms of bias and variance of the policy gradient estimation? Explain your answer.

Question 1.4 — 10 points How could the bias-variance tradeoff in actor-critic be controlled?

Aalto University
School of Electrical
Engineering

Reinforcement Learning course staff
Aalto Robot Learning Lab
aalto.fi, rl.aalto.fi

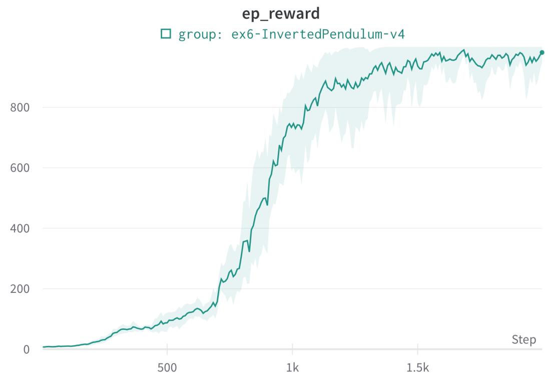

Figure 1: Training plot of the policy gradient with a critic.

Deep deterministic policy gradient

Task 2 — 25 points Implement the deep deterministic policy gradient (DDPG) algorithm for the HalfCheetah environment. The code is in the `ddpg.py`. Code from `ex4/dqn.py` might be helpful. If needed, you can also reference the paper [1]. **Attach the training performance plot into your report.**

The reference training plot is as Figure 2:

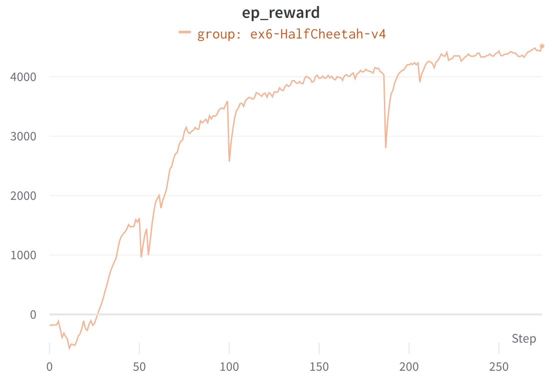

Figure 2: Training plot of the deterministic policy gradient on the HalfCheetah environment.

Question 2.1 — 10 points For policy gradient methods seen in Exercise 5, we update the agent using only on-policy data, while in DDPG we can use off-policy data. Why is this the case?

Question 2.2 — 10 points A big advantage of DDPG is that it's able to utilise off-policy data. What are the disadvantages of deterministic policy gradient compared to the policy gradient method implemented in Task 1? List two of them.

Submission

The deadline to submit the solutions through MyCourses is on Monday, 07.11 at 23:55.

1. **Answers to all questions** posed in the text.
2. The **training performance plots** for each of the tasks (Tasks 1 and 2).

In addition to the report, you must submit as separate files, in the same folder as the report:

1. Python **code** used to solve **all task exercises**.

Please remember that not submitting a PDF report following the **Latex template** provided by us will lead to subtraction of points.

For more formatting guidelines and general tips please refer to the submission instructions file on mycourses.

If you need help or clarification solving the exercises, you are welcome to join the exercise sessions.

Good luck!

References

[1] Timothy P. Lillicrap et al. *"Continuous control with deep reinforcement learning"* ICLR 2016.